Watching a drunkard for ten nights: A study of distributions of variances

R. K. P. Zia and B. Schmittmann
Center for Stochastic Processes in Science and Engineering,
Physics Department, Virginia Polytechnic Institute and State University,
Blacksburg, VA, 24061-0435 USA

(February 1, 2008)

For any physical observable in statistical systems, the most frequently studied quantities are its average and standard deviation. Yet, its full *distribution* often carries extremely interesting information and can be invoked to put any surprising properties of the individual moments into perspective. As an example, we consider a problem concerning simple random walks which was posed in a recent text. When a drunk is observed over L nights, taking N steps per night, and the number of steps to the right is recorded for each night, an average and a variance based on these data can be computed. When the variance is used to estimate p, the probability for the drunk to step right, *complex* values for p are frequently found. To put such obviously nonsensical results into context, we study the full probability distribution for the variance of the data string. We discuss the connection of our results to the problem of data binning and provide two other brief examples to demonstrate the importance of full distributions.

I. INTRODUCTION

Many properties of the random walk are well-known[1] and form an important part of good texts in statistical mechanics, at either the undergraduate or graduate level[2]. Despite its "age,"[3] this problem continues to present new and interesting puzzles, depending on the questions asked of the walker. Many recent examples of such puzzles, some of which remain unsolved, involve the issue of *full distributions* (as opposed to just the averages) of certain quantities. Providing an exhaustive list is beyond the scope of this paper. Instead, in the concluding section, we discuss several examples to illustrate both the value and the excitement in such studies. Our main interest lies in a distribution rarely discussed in texts, in connection with a problem posed in the manuscript of an undergraduate text book on statistical mechanics, co-authored by H. Gould and J. Tobochnik (GT)[4].

Let us first quote the question which motivated this paper, namely, the first two parts of problem 3.37 in this manuscript.

> A random walker is observed to take a total of N steps, n of which are to the right.
>
> *(a)* Suppose that a curious observer finds that on ten successive nights the walker takes $N = 20$ steps and that the values of n are given successively by 14, 13, 11, 12, 11, 12, 16, 16, 14, 8. Compute $\bar{n}$, $\overline{n^2}$, and σ_n. Use this information to estimate p. If your reasoning gives different values for p, which estimate is likely to be the most accurate?
>
> *(b)* Suppose that on another ten successive nights the same walker takes $N = 100$ steps and that the values of n are given by 58, 69, 71, 58, 63, 53, 64, 66, 65, 50. Compute the same quantities as in part (a) and estimate p. ... Explain your results.

To help the reader with context and notation, we add that this problem is at the end of a section about a simple random walk in one dimension, stepping either to the right or left, with probability p and $1-p$, respectively. The standard results for the average number of right steps, Np, and the associated variance, $Np\,(1-p)$, were derived. We will refer to these expressions as the "true average" and the "true variance," denoted by

$$n_{av} \equiv Np \quad \text{and} \quad v \equiv Np\,(1-p) \tag{1}$$

respectively. They would result if we were to observe the walker for very (ideally, infinitely) many nights.

In the problem, however, $\bar{n}$ and σ_n denote the average and the standard deviation (i.e., $\sqrt{\overline{n^2} - \bar{n}^2}$) computed from observations covering only a relatively *small* number (ten) of nights. So, estimates for p are to be made from either the equation for the average

$$\bar{n} = Np_{av} \tag{2}$$

or one for the standard deviation

$$\sigma_n = \sqrt{Np_\sigma\,(1-p_\sigma)}\ . \tag{3}$$

The authors pointed out[5] that, since the second method involves a quadratic equation for p, there will be *two* solutions (indeed, symmetric around $p = 1/2$), so that this route cannot produce a unique answer by itself.

More interestingly, if straightforward computations are carried out, it is even more trivial to answer the question concerning which estimate is more "accurate": p_σ is *complex* for both (a) and (b)! When we trace the origins of this remarkable result, we find that, though the true variance v never exceeds $N/4$, the data string of a *particular* night can easily exceed this bound. One of the "worst case" scenarios – focusing again on 10 nights – occurs when the drunk takes N steps to the right for $10p$ nights and N steps to the left for $10\,(1-p)$ nights. For example, instead of the first string given in part (a) of the problem, we have (assuming $p = 0.6$) 20,20,0,0,20,0,20,20,0,20. Though such a string even provides the exact underlying p, it leads to $\sigma_{\max}^2 = N^2 p(1-p)$. Note that $\sigma_{\max}^2$ comes with even a *wrong power* of N, so that, for sufficiently large N, it will always exceed the largest possible true variance, $\max_{p\in[0,1]} Np(1-p) = N/4$, regardless of the underlying p! At the other extreme, if the same n is observed on *every* night, then the absolute minimum $\sigma_{\min} = 0$ is achieved. Though this $\sigma_{\min}$ leads to a *real* p_σ, the result ($p_\sigma = 0$ or 1) is clearly also unreasonable. Although we can easily compute the expectation values of σ_n and p_σ, a natural question arises: how likely is it that σ_n^2 exceeds the absolute bound of $N/4$? Similarly, we could ask for the likelihood that p_σ comes within, say, 5% of the underlying p. Thus, we are led to study the full distribution of variances in the observations of a drunkard.

The rest of this paper is organized as follows. The next section is devoted to the detailed analysis of the issues at hand. Section III is provided as a convenience, for those who wish to skip the details for now, read a summary of our conclusions, and see how they apply to the specific problem here. In a final section, we make brief suggestions for expanding discussions of distributions in typical courses on statistical mechanics.

II. DISTRIBUTIONS OF AVERAGES AND VARIANCES

To motivate the study of distributions, let us use the language of human behaviorists to define an "ensemble" of many *identical* drunkards (say, M) residing in different cities. The probability of stepping to the right is p, for *all* drunkards in *all* nights. Each is observed for L nights, N total steps each night. The number of steps to the right, n, differs each night and for different walkers. So, the entire data set can be summarized by M strings of L numbers:

$$n_{\alpha,i} \ , \quad \alpha \in [1, M]\,;\, i \in [1, L] \ , \tag{4}$$

with

$$0 \le n_{\alpha,i} \le N \ .$$

Alternatively, we can imagine the string of numbers in GT's text being generated randomly (with the *same* p) each time the URL is accessed. After M readings, we would have M different strings. So, for the two cases stated in the problem above, M is just 1 and $L = 10$, while $N = 20$ and 100.

From this big data set, we can construct M averages and variances, each generated from the data set for one of the drunks:

$$\bar{n}_\alpha \equiv \frac{1}{L}\sum_i n_{\alpha,i} \,,\ V_\alpha \equiv \frac{1}{L}\sum_i n_{\alpha,i}^2 - (\bar{n}_\alpha)^2 \ . \tag{5}$$

Needless to say, $0 \le \bar{n}_\alpha \le N$ and $0 \le V_\alpha \le N^2/4$. (Note the N^2 in the last bound, which comes from *the worst* case scenario with only right steps for exactly half the nights!) Clearly, $\bar{n}_\alpha$ and V_α are still random variables, with respect to the ensemble of drunks. If we make histograms from the two sets of M numbers, denoted by $H(\bar{n})$ and $H(V)$, we get a glimpse of the full distribution of their possible values. The first is easy, being related to the binomial distribution, and approaches a Gaussian for large N. Nevertheless, there are some subtleties to which we will alert the reader. The second will give us an idea of how often we can expect a data string to produce a variance which exceeds the maximum possible value, for any realizable p, namely, $N/4$.

A. Statistics of averages

For completeness, let us remind the reader that, if we focus on the statistics of a single drunkard taking N steps on a single night, then the probability that he takes n right steps is just the binomial distribution:

$$P(n) = \binom{N}{n} p^n (1-p)^{N-n} \ . \tag{6}$$

If he is observed for L nights (for a total of LN steps), then the probability that he took a *total* of m steps to the right is clearly also a binomial: $\binom{LN}{m} p^m (1-p)^{LN-m}$. But, notice that $\bar{n}$, the average over the L nights of right steps, is precisely m/L. Thus, we may conclude immediately that the probability distribution for the *average* number of right steps is

$$\mathcal{P}(\bar{n}) = \binom{LN}{L\bar{n}} p^{L\bar{n}} (1-p)^{L(N-\bar{n})} \ . \tag{7}$$

If we take data for M drunkards and compile a histogram for the set of $\bar{n}_\alpha$, we should find

$$H(\bar{n}) \to M\mathcal{P}(\bar{n}) \tag{8}$$

as $M \to \infty$.

This distribution also tells us how reliable our estimate for p is when we have data for only one drunkard (as posed in the text problem). In particular, as we can guess intuitively, although we are *most likely* to get the right p by just dividing $\bar{n}$ by N (i.e., $\mathcal{P}(\bar{n})$ peaks at Np), the chances we are off can be estimated through the standard deviation in $\mathcal{P}(\bar{n})$, namely, $Np(1-p)/L$.

Finally, if N is large, the binomial $P(n)$ is indistinguishable from a Gaussian:

$$\tilde{P}(n) = \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{(n-n_{av})^2}{2v}\right] \ . \tag{9}$$

Since convolutions of Gaussians form a Gaussian, the distribution of the averages is also a Gaussian, with variance

$v/L.$ In other words, we should find

$$\frac{H(\bar{n})}{M}\underset{M\to\infty}{\to}\tilde{\mathcal{P}}(\bar{n};n_{av},v,L)=\sqrt{\frac{L}{2\pi v}}\;e^{-\frac{L(\bar{n}-n_{av})^2}{2v}} \quad (10)$$

Note that we have included an explicit list of the parameters which control the distribution $\tilde{\mathcal{P}}$.

B. Statistics of variances

Next, we turn to the central issue of this paper. If p is estimated through the variance of a single string of L observations (n_i), how often can we expect the estimate to be complex? In other words, what is the probability that the variance $L^{-1}\sum_i n_i^2-\bar{n}^2$ exceeds the absolute bound

$$V_{abs}\equiv N/4\;? \quad (11)$$

To answer this question, we focus on the probability that $L^{-1}\sum_i n_i^2-\bar{n}^2$ assumes a given value, say, V. First, we ask for the (joint) probability that the number of right steps, observed over L nights, takes the values n_1, n_2,..., n_L. Since the events of any night are independent of all other nights, this joint probability $P(n_1,n_2,...,n_L)$ is just the product of the probabilities for a single night, i.e., $\prod_{\ell=1}^{L}P(n_\ell)$ – and $P(n_\ell)$ is simply given by the binomial, Eqn (6). Summing over all possible outcomes $\{n_\ell\}\equiv\{n_1,n_2,...,n_L\}$ by using a Kronecker delta to count only those for which the variance equals V, we obtain

$$\mathcal{P}(V)=\sum_{\{n_\ell\}}\delta\left(V,\frac{1}{L}\sum_{i=1}^{L}n_i^2-\left[\frac{1}{L}\sum_{i=1}^{L}n_i\right]^2\right)\prod_{\ell=1}^{L}P(n_\ell)\;. \quad (12)$$

Unfortunately, we are unable to evaluate this expression exactly. However, an excellent approximation can be obtained if N is not too small (as in the cases posed in Ref. 4) so that we can use the Gaussian approximation, Eqn (9), instead of the exact binomial, Eqn (6). A reader intimately familiar with error analysis will recognize our quest as the probability density function for the χ^2-distribution[6]. For pedagogical purposes, let us show how to make progress from this point. Let the probability that the variance lies in the interval $[V,V+dV]$ be $\tilde{\mathcal{P}}(V)\,dV$. Then,

$$\begin{aligned}\tilde{\mathcal{P}}(V;v,L)&=\prod_{\ell=1}^{L}\int_{-\infty}^{\infty}dn_\ell\tilde{P}(n_\ell)\times\\ &\delta\left(V-\frac{1}{L}\sum_{\ell=1}^{L}n_\ell^2+\left[\frac{1}{L}\sum_{\ell=1}^{L}n_\ell\right]^2\right)\;, \quad (13)\end{aligned}$$

where we have included the relevant parameters (v,L) explicitly. Note that all variables are now continuous with infinite range and, correspondingly, the δ here is the Dirac delta. To evaluate this distribution, consider its Laplace transform

$$\begin{aligned}\mathcal{L}(\mu)&\equiv\int dVe^{-\mu V}\tilde{\mathcal{P}}(V;v,L)\\ &=\prod_{\ell=1}^{L}\int_{-\infty}^{\infty}dn_\ell\tilde{P}(n_\ell)\times\\ &\exp\left[-\frac{\mu}{L}\sum_{\ell=1}^{L}n_\ell^2+\frac{\mu}{L^2}\left(\sum_{\ell=1}^{L}n_\ell\right)^2\right]\;. \quad (14)\end{aligned}$$

Since each $\tilde{P}$ is a Gaussian, we have here a generalized Gaussian integral. Explicitly, we have

$$\mathcal{L}(\mu)=\prod_{\ell=1}^{L}\int_{-\infty}^{\infty}\frac{dn_\ell}{\sqrt{2\pi v}}\times \quad (15)$$
$$\exp\left[-\frac{1}{2v}\sum_{\ell=1}^{L}(n_\ell-Np)^2-\frac{\mu}{L}\sum_{\ell=1}^{L}n_\ell^2+\frac{\mu}{L^2}\left(\sum_{\ell=1}^{L}n_\ell\right)^2\right]$$

To proceed, it is best to displace each n_ℓ by Np. Verifying that the variance is independent of such a shift, we have a simpler form:

$$\begin{aligned}\mathcal{L}(\mu)&=\prod_{\ell=1}^{L}\int_{-\infty}^{\infty}\frac{dn_\ell}{\sqrt{2\pi v}}\times \quad (16)\\ &\exp\left[-\left(\frac{1}{2v}+\frac{\mu}{L}\right)\sum_{\ell=1}^{L}n_\ell^2+\frac{\mu}{L^2}\left(\sum_{\ell=1}^{L}n_\ell\right)^2\right]\end{aligned}$$

Rescaling each n_ℓ by $\sqrt{vL/(L+2\mu v)}$ simplifies this expression further:

$$\begin{aligned}\mathcal{L}(\mu)&=\left(\frac{L}{L+2\mu v}\right)^{L/2}\prod_{\ell=1}^{L}\int_{-\infty}^{\infty}\frac{dn_\ell}{\sqrt{2\pi}}\times \quad (17)\\ &\exp\left\{-\frac{1}{2}\left[\sum_{\ell=1}^{L}n_\ell^2-\frac{2\mu v}{(L+2\mu v)}\left(\frac{1}{\sqrt{L}}\sum_{\ell=1}^{L}n_\ell\right)^2\right]\right\}\end{aligned}$$

The exponent should be regarded as a quadratic form

$$-\frac{1}{2}\sum n_\ell Q_{\ell\ell'}n_{\ell'} \quad (18)$$

where the matrix is just an identity matrix plus a term proportional to the tensor product of a single unit vector, namely, $(1,1,...,1)/\sqrt{L}$. The eigenvalues of such a matrix are all unity, except for one, which is unity plus the value of this proportionality constant. In Eqn (17), we have purposefully written the last term in the exponent to display this constant, i.e., $-2\mu v/(L+2\mu v)$. Exploiting

$$\prod_{\ell=1}^{L}\int_{-\infty}^{\infty}\frac{dn_\ell}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\sum n_\ell Q_{\ell\ell'}n_{\ell'}\right]=(\det Q)^{-1/2}\;, \quad (19)$$

we have

$$\mathcal{L}(\mu) = \left(\frac{L}{L+2\mu v}\right)^{L/2}\left(1-\frac{2\mu v}{(L+2\mu v)}\right)^{-1/2} = \left(\frac{L}{L+2\mu v}\right)^{(L-1)/2} . \quad (20)$$

To obtain $\tilde{\mathcal{P}}(V;v,L)$, we only need to perform an inverse Laplace transform. Fortunately, $\mathcal{L}(\mu)$ is simple enough to appear in standard tables[7] and the explicit answer is

$$\tilde{\mathcal{P}}(V;v,L) = \frac{1}{V\Gamma\left(\frac{L-1}{2}\right)}\left(\frac{LV}{2v}\right)^{(L-1)/2} e^{-LV/2v} . \quad (21)$$

Of course, this distribution may be written in scaling form:

$$\tilde{\mathcal{P}}(V;v,L) = \frac{L}{2v}\Phi_{\frac{L-1}{2}}(x) \quad (22)$$

where the (non-negative) scaling variable is

$$x \equiv \frac{LV}{2v} , \quad (23)$$

and Φ is the standard gamma distribution[6]:

$$\Phi_\gamma(x) = \frac{x^{\gamma-1}}{\Gamma(\gamma)}e^{-x}. \quad (24)$$

with

$$\gamma \equiv (L-1)/2$$

for our case. We briefly note some of its properties: It peaks at $x_{peak} = \max\{0, \gamma-1\}$ and has moments $\langle x^n\rangle = \Gamma(n+\gamma)/\Gamma(\gamma)$. Thus, its average and variance both equal γ: $\langle x\rangle = \langle x^2\rangle - \langle x\rangle^2 = \gamma$.

Returning to the problem we posed for the human behaviorist, namely, to compile a histogram for the set of V_α, Eqn (5), on the basis of data for M drunkards, Eqn (4), we should find

$$H(V) \to M\tilde{\mathcal{P}}(V;v,L) \quad (25)$$

for very large M.

Before applying this result to the problem quoted above, let us comment on several interesting features.

From Eqn (23), we might infer that the relevant scale for V is $v/L = Np(1-p)/L$. Nonetheless, the *average* (or expectation value) of V , being

$$V_{av} = \frac{2v}{L}\gamma = Np(1-p)\left[1-\frac{1}{L}\right] , \quad (26)$$

is much closer to v, especially for large L. Another interesting quantity is the *most likely* value for V, which is simply related to the peak value of Φ_γ, via $V_{peak} = (2v/L)\,x_{peak}$. So,

$$V_{peak} = \begin{cases} 0 & \text{for } L \le 3 \\ Np(1-p)\left[1-\frac{3}{L}\right] & \text{for } L > 3 \end{cases} . \quad (27)$$

Both are consistently less than $Np(1-p)$, the "true variance". Though both approach v monotonically as L increases, these results provide a measure of how much the variance of a short data string (i.e., small L) can differ from its asymptotic ($L\to\infty$) value.

Since short data strings lead to the most serious discrepancies, let us make a detour for the cases $L = 1, 2$, and 3. The appearance of $\Gamma\left(\frac{L-1}{2}\right)$ in the denominator of Eqn (21) ensures that, for $L = 1$, $\tilde{\mathcal{P}} \equiv 0$ for $V \neq 0$. On the other hand, this distribution is normalized, so that we may conclude $\tilde{\mathcal{P}}(V;v,0) = \delta(V)$, regardless of v. This result is perfectly understandable, since $L = 1$ corresponds to our observing the drunkard for only a *single* night. If a "data set" consists of only a *single* number, the "variance" is necessarily zero, regardless of the behavior of the drunkard! For $L = 2$, $\tilde{\mathcal{P}}$ diverges at the origin. There is no cause for alarm, however, as the divergence is weak enough for $\int \tilde{\mathcal{P}}dV$ to be finite in any neighborhood of $V = 0$. Indeed, a better perspective is provided by the distribution for the standard deviation

$$\sigma \equiv \sqrt{V}$$

Then, we find a familiar looking expression:

$$P(\sigma;v,2) \equiv \frac{dV}{d\sigma}\tilde{\mathcal{P}}(V;v,2) = \frac{2}{\sqrt{\pi v}}e^{-\sigma^2/v}.$$

However, keep in mind that we have only "half a Gaussian" here, since $\sigma \in [0,\infty]$! Finally, the other curious case is $L = 3$ where $\mathcal{P}(V;v,3)$ is a pure exponential, for which the standard deviation assumes the same value as the average. By contrast, the situation for large L provides few surprises. The gamma distribution approaches a Gaussian with width $O\left(1/\sqrt{L}\right)$.

III. SUMMARY AND APPLICATION TO THE GOULD-TOBOCHNIK[4] PROBLEM

Let us summarize our findings before applying them to the problem quoted in the Introduction. For the reader's convenience, we reiterate the set-up: Consider an "ensemble" of M *identical* drunkards with p as the probability of stepping to the right. Each is observed for L nights, taking N total steps each night. For drunkard α $(= 1, ..., M)$, the number of right steps is recorded for each night ($i = 1, ...L$) and denoted by $n_{\alpha,i}$. From these, we compute the averages $\bar{n}_\alpha \equiv L^{-1}\sum_i n_{\alpha,i}$ and variances $V_\alpha \equiv L^{-1}\sum_i n_{\alpha,i}^2 - (\bar{n}_\alpha)^2$. Normalized histograms for these are compiled and denoted by $H(\bar{n})/M$ and $H(V)/M$. We find that, already for moderate N (such as 20), the first will approach a Gaussian distribution while the second, a gamma distribution. Explicitly,

$$\frac{H(\bar{n})}{M} \underset{M\to\infty}{\to} \tilde{\mathcal{P}}(\bar{n};n_{av},v,L) = \sqrt{\frac{L}{2\pi v}}\, e^{-\frac{L(\bar{n}-n_{av})^2}{2v}} \quad (28)$$

and

$$\frac{H(V)}{M} \underset{M\to\infty}{\to} \tilde{\mathcal{P}}(V; v, L) = \frac{x^{\gamma} e^{-x}}{V\Gamma(\gamma)} \ . \qquad (29)$$

where $n_{av} \equiv Np$, $v \equiv Np(1-p)$, $x \equiv LV/2v$, and $\gamma \equiv (L-1)/2$.

Applying these results to the specific problem at hand, let us *assume* that the underlying p is 0.6. With $L = 10$ and $N = 20$ or 100, we have the full distribution of variances

$$\tilde{\mathcal{P}}(V; 0.24N, 10) = \frac{x^{4.5} e^{-x}}{V\Gamma(4.5)}. \qquad (30)$$

with

$$x = \frac{V}{0.048N}. \qquad (31)$$

As noted in the previous section, the peak of $\tilde{\mathcal{P}}$ occurs at $v[1 - 3/L]$, which is

$$V_{peak} \cong 0.168N \qquad (32)$$

in this case. Although this value appears far below the maximum allowed, $V_{abs} = 0.25N$, we should be more careful with our original question:

How likely will V exceed V_{abs}?

To answer this question, we consider the integral

$$\rho(p, L) \equiv prob(V > V_{abs}) = \int_{N/4}^{\infty} \tilde{\mathcal{P}}(V; v, L)\, dV \ , \quad (33)$$

Using the scaled variable, ρ is just

$$\rho(p, L) = \frac{1}{\Gamma(\gamma)} \int_{x_{abs}}^{\infty} x^{\gamma-1} e^{-x} dx \qquad (34)$$

which should be recognized as a standard χ^2-probability function[6] evaluated at a particular point:

$$x_{abs} = \frac{LV_{abs}}{2Np(1-p)} = \frac{\gamma + 1/2}{4p(1-p)} \ .$$

Note that ρ is now manifestly *independent* of N, the number of steps taken by each drunk per night. As a result, we have written explicitly in Eqn (34) that it depends solely on p and L.

Applying this integral to the case at hand, we write

$$\rho(0.6, 10) \cong \frac{1}{\Gamma(4.5)} \int_{5.2083}^{\infty} x^{3.5} e^{-x} dx \qquad (35)$$

and, computing the integral numerically, find

$$\rho(0.6, 10) \cong 0.3178 \ . \qquad (36)$$

This is an astonishingly large value! In other words, by observing M drunkards (identical ones, with 0.6 probability of stepping to the right) for 10 nights each, about $M/3$ of the data sets will lead to a variance exceeding the absolute bound of $N/4$! In the equivalent scenario, if the string of numbers in the text[4] were generated anew with each access to the URL, then about *a third* of the students would find complex p_σ's by using σ_n, Eqn (3). Let us remind the reader that this conclusion is *independent* of N!

For students to appreciate these considerations better, it suffices to set up a crude Excel spreadsheet. Assuming $p = 0.6$ and placing the formula IF(RAND()¡0.6,1,0) in, e.g., the block A1-T12000, we can "create" the data (with right/left steps recorded as 1/0) for $M = 1200$ drunkards, observed on $L = 10$ nights (10 successive rows for each drunkard), for $N = 20$ steps (columns A-T). Summing each row into column U, we have the "data set" $\{n_{\alpha,i}\}$. After computing the averages and variances for the numbers in column U (in blocks of 10), the histogram function can be invoked to provide $H(\bar{n})$ and $H(V)$. In Fig. 1, we show the result for the latter (for a particular run, of course). Since the peak position is expected to be at 3.36 (considerably less than the "true variance" of 4.8!), we choose bins of width 0.4, so that the center of one bin (3.4) approximately coincides with the peak. As we see from the histogram, the peak frequency indeed lies in the appropriate bin. In Fig. 1, we also plot the theoretical distribution, Eqn (30), multiplied by $M = 1200$. The zero-parameter "fit" is clearly excellent. Since the peak is quite far from the average, we have also computed the latter and find $V_{av} \cong 4.331$. This value is entirely consistent with the predicted $(0.24)(20)(0.9) = 4.32$, its approximate location being indicated by an arrow in Fig. 1. Lastly, we program the spread sheet to find the number of V's lying above the bound of 5 in this sample. The result, for this particular run was 374, again consistent with the predicted $381 = (0.3178)(1200)$.

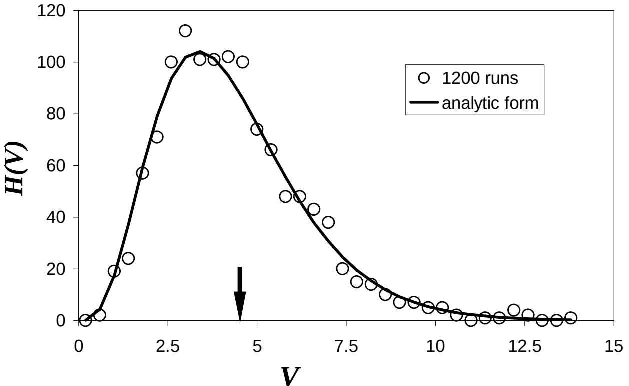

FIG. 1. Histogram and analytic form for the distribution of variances. The average variance, $V_{av} \simeq 4.3$, is indicated by the arrow. The data are "observations" of 1200 drunkards for 10 nights, each taking 20 steps each night.

Given this large percentage of obviously non-sensical estimates, a conscientious observer would naturally attempt to improve the situation, by making more observations. Noting that this difficulty is *independent* of N

(the number of steps tallied on each night), our observer might try increasing L, by watching the drunks for more than 10 nights. One might hope that taking data for 100 nights, say, should surely lead to a much larger proportion of "good data" (i.e., variances that lead to real estimates for p). Mustering extreme patience, our observer collects the necessary data, performs the analysis and is absolutely shocked to discover the result: instead of dropping sharply, ρ has *increased even further*, to 0.3416! In fact, a little thought shows that ρ cannot be a monotonically decreasing function of L. Indeed, for $L = 1$, $\rho\,(p, 1)$ *vanishes*, since $\tilde{\mathcal{P}} = \delta\,(V)$. So, at least initially, $\rho\,(0.6, L)$ increases with L. In fact, it peaks at $L = 41$ where it takes its maximum value of 0.35553. Beyond $L = 41$, it decreases, but so slowly that $\rho\left(0.6, 10^2\right) \cong 0.3416$ is still larger than $\rho\,(0.6, 10) \cong 0.3178$. Indeed, it decreases exceedingly slowly, so that, e.g., $\rho\left(0.6, 10^3\right) \cong 17\%$!

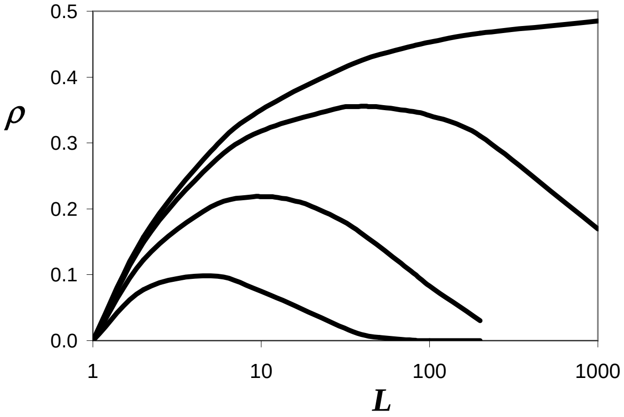

FIG. 2. The probability $\rho\,(p, L)$ that the variance will exceed the absolute bound $N/4$, plotted vs L for various p. The curves, from the top, are for $p = 0.5, 0.6, 0.7$, and 0.8, respectively.

Since ρ is identically zero at $L = 1$ (for any p), while $\rho\,(p, L \to \infty)$ is expected to vanish, an interesting question is the following. Given p, for what value of L will ρ reach its maximum? Though easy to pose, we have not found a simple analytic answer to this question. Instead, in Fig. 2, we present plots of $\rho\,(p, L)$ for a few p's, so that our readers can appreciate its general behavior. Let us end by simply stating another remarkable result: for the *unbiased* walk, $\rho\,(0.5, L)$ is a *monotonically increasing* function of L, with limiting value of $1/2$. In other words, as the number of observation nights is increased, the likelyhood of the computed variance exceeding the true one *increases*. In the limit of compiling data from "infinitely many" nights (of N-step walks), it is equally likely for the variance to exceed $N/4$ (the true value) as otherwise! For students interested in critical phenomena, it is possible to regard $(p, L) = (0.5, \infty)$ as a critical point and to seek a scaling function for ρ involving $(p - 0.5)\,L^{1/4}$. A full analysis of these issues would be more suitable, we believe, for the readership of another journal.

IV. CONCLUDING REMARKS

We have analyzed an interesting problem posed in the manuscript of Gould and Tobochnik[4]. A random walker is observed for 10 nights, taking 20 steps each night. The number of *right* steps taken each night is recorded, and the student is asked to compute the average and variance from, and for, these 10 data points. Letting p denote the probability of the walker taking a right step, the true variance is just $20p(1 - p)$ and can never exceed 5. Nevertheless, this bound *is exceeded* for the specific sequence of ten numbers in the text! In an effort to resolve this surprisingly result, we find that, if many such sequences (of ten numbers) were generated according to $p = 0.6$, nearly 32% of them will lead to a variance exceeding the absolute bound. Indeed, we found the full distribution for the variance, $\tilde{\mathcal{P}}\,(V; L, N, p)$, for the general case ($L$ nights, N steps, arbitrary p).

There is an alternative perspective to the behavior discussed here, namely, in terms of the statistics associated with "binning." After all, if each step of the random walker is independent, then – instead of binning the data into L nights of N steps each – we might as well consider the whole data set as a single string of NL steps. Associated with this string is, of course, just a single number: $\tilde{n}$, the total number of right steps. Seeking, as before, the "average" and "variance" of this long string, the "average" is simply $\tilde{n}$, giving just *one* estimate for p, i.e., $\tilde{n}/NL$. Moreover, the "variance" is necessarily zero. Now, by observing M drunkards, we can make a histogram of the associated estimates for p. For large M, this histogram will approach the theoretical distribution of a binomial associated with $\binom{NL}{\tilde{n}}$. If NL is also large, the histogram is well approximated by a Gaussian centered on p with standard deviation $\sqrt{p\,(1-p)\,/NL}$. To write this string in the setting of the textbook problem, we would *bin* the NL steps into L bins and arrive at L numbers n_i, as well as L estimates (for p): n_i/L. The average of these estimates is *exactly* $\tilde{n}/NL$, i.e., as if no binning took place. However, by binning, we have created a non-trivial "variance," associated with the binned values n_i! Since it will depend on the bin size (or equivalently, the number of bins: L), we should denote such a variance by V_L. Above, we have presented its distribution and its non-trivial dependence on L.

Our goal in this article is to demonstrate that the full *distribution* of a physical quantity often carries extremely interesting information and can be invoked to put surprising properties of individual moments into perspective. We believe that this message is central to statistical mechanics and should be taught within the context of upper level courses. Of course, the notion of full distributions *is* frequently included in texts but typically limited to the binomial and Gaussian distributions. In the context of the random walk and related problems, these lead naturally to a discussion of the central limit theorem. This theorem elevates the Gaussian up to the status of a "universal" distribution. The appeal of this notion is so strong that students might be lulled into thinking that *all* distributions are "normal" or "bell shaped," for which only average and variance are needed. Yet, many distributions are not normal and display a variety of interesting properties. A good example is the Poisson distribution, included in GT[4] to convince readers that they need not be afraid of flying. For these reasons, we believe that it would be valuable to devote some lessons in a course on statistical mechanics, or an entire section of a graduate or an undergraduate text, to distributions. In particular, the notion of derived, or induced distributions is very important: In many areas of physics, we are frequently interested in stochastic quantities which are themselves *specific functions* of other, underlying random variables (often assumed to be Gaussian). Of course, finding the full distribution of such quantities can turn into a very challenging task but the rewards can be considerable, as we seek to demonstrate with two examples, both of which are relatives of the simple one-dimensional random walk. The first, concerning the statistical widths of a one-dimensional interface leads to a *universal* (but non-Gaussian) distribution, describing a *whole class* of interface models. The second illustrates the observation that, depending on the question posed, even simple random walks can give rise to non-analyticities which may be difficult to explore fully.

Statistical widths of a one-dimensional interface, or average deviations in a random walk. The simple random walk on a line is just a string of R and L steps. Plotting the displacement on the y-axis and the number i of steps taken on the abscissa, a particular walk can be viewed as a specific configuration of a *one-dimensional* interface, embedded in a two-dimensional bulk: the interface *height* h_i is measured along y vs the "column label" $i = 1, 2, ...N$. In the language of interfaces, two natural and frequently considered quantities are the average height $\bar{h} \equiv \sum_i h_i/N$ and width (easily recognized as a standard deviation) $w \equiv \sqrt{\sum_i \left(h_i - \bar{h}\right)^2/N}$ of a given interface. However, in the context of random walks, neither of these quantities comes to mind easily. Instead of being an ensemble average over very many walks, $\bar{h}$ is the average displacement over *time*, of a *specific* walk in a *specific* night, and hence depends on the entire history of this particular walk. Similarly, w is different from one interface to another and, in terms of the drunkard, typically changes from one night to the next. Here, we are interested in the full distributions of $\bar{h}$ and w. Obviously, we can compile their histograms, $H\left(\bar{h}\right)$ and $H(w)$, by generating many interface configurations or by observing the drunkard over many nights. If the underlying probability is symmetric ($p(\pm 1) = 1/2$), then it is easy to imagine that $H\left(\bar{h}\right)$ is symmetric and approaches a Gaussian, thanks to the central limit theorem. On the other hand, it is not so easy to guess what $H(w)$ should be. First, w is never negative, so $H(w)$ must "end" at $w = 0$. Second, though the average (and perhaps the peak position) should increase with $\sqrt{N}$, the "top end" of $H(w)$ will be $O(N)$. Referring to Ref. 8 for details, we just summarize the key results. Noting that $\left\langle w^2\right\rangle$, the width-square *averaged* over all possible configurations, is just $L/12$, we obtain $H(w) \sim \exp\left[-\pi^2 w^2/\left\langle w^2\right\rangle\right]$ for large widths and $H(w) \sim w^{-5}\exp\left[-3\left\langle w^2\right\rangle/2w^2\right]$ for small w's. In other words, configurations with widths significantly larger/smaller than $\left\langle w^2\right\rangle$ are exponentially suppressed. Remarkably, this distribution is *universal*, describing a whole class of interface models, in the same sense that the Gaussian is a limiting distribution for sums of random numbers.

Distribution of longest returns. Here, we turn to an extremely interesting example, in which there is no analytic solution even for simple one-dimensional random walks. Motivated by the physics of charged polymers (a string of monomers, each carrying charge ±1), Ertas and Kantor[9] explored the properties of the largest neutral segments. Translated into the language of one-dimensional random walks, a neutral segment, consisting of equal numbers of opposite charges, corresponds to a part of the walk where the walker returns to a particular site. So, the "largest neutral segment" maps into the "longest return path," or largest number of steps between visits to the same site (regardless of which site). A more explicit phrasing of the question is: "Of all the 2^N random walks of N steps on a line, how many have N' (with $0 \le N' \le N$) as the longest return path?" Denoting the answer by $H(N'; N)$, the normalized distribution is $P(N') = H(N'; N)/2^N$. Again, it may be helpful to consider the extremes. For $N' = 0$, there are precisely two walks: all right or all left steps. So, $H(0; N) = 2$. At the other extreme, for $N' = N$, these walks are the familiar ones which return to the starting point, so that $H(N; N) = \binom{N}{N/2}$. In the limit of large N, it is better to use the fraction $\phi \equiv N'/N$ as a variable and to take the continuum limit, so that an appropriate probability density, $p(\phi)$, emerges from $P(N')$. Remarkably, $p(\phi)$ develops a kink, i.e., a *discontinuous* first derivative, at $\phi = 1/2$ (Fig. 3 in Ref. 9)! Despite the simple sounding nature of this question, it is clearly quite a complex issue. Though some understanding of this unexpected phenomenon is possible, it is far from being completely solved.

To conclude, we hope to have motivated teachers and students alike to think beyond the first few moments and to focus on full distributions, if at all possible. To whet our readers' appetites, we have presented a particularly striking example in the context of simple random walks: if one uses finite data strings to estimate the (asymptotic) probability p, the estimate can easily turn out to be manifestly nonsensical, namely, complex! Beyond this "demonstration", we briefly discussed two further examples, all associated with random walks. Hopefully, these concepts will challenge our readers to explore, or discover, their own favorite distributions.

Acknowledgements. We thank Harvey Gould, Jan Tobochnik, and Uwe Täuber for very helpful discussions. This work is supported in part by a grant from the National Science Foundation, through the Division of Materials Research.

[1] B.D. Hughes, *Random Walks and Random Environments*, Vol.1 (Clarendon Press, Oxford 1995)

[2] See, e.g., F. Reif, *Fundamentals of Statistical and Thermal Physics* (McGraw Hill, New York 1965) and[4] below.

[3] K. Pearson, Letter to Nature, Nature **72**, 294 (1905); Lord Rayleigh, Phil. Mag. **10**, 73 (1880); M. von Smoluchowski, Ann. Phys. **21**, 756 (1906); G. Pólya, Mitteilungen der Physikalischen Gesellschaft Zürich **19**, 75 (1919). A brief history of the random walk problem can be found in[1].

[4] H. Gould and J. Tobochnik, *Thermal and Statistical Physics*, available on-line: http://stp.clarku.edu/gtdraft/.

[5] H. Gould and J. Tobochnik, private communication (2002).

[6] See, e.g., NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/eda /section3/eda3666.htm, 2002

[7] See, e.g., M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, (Dover Publications, New York 1974)

[8] G. Foltin, K. Oerding, Z. Rácz, R.L. Workman, and R.K.P. Zia, Phys. Rev. E **50** R639 (1994).

[9] D. Ertas and Y. Kantor, Phys. Rev. **E53**, 846 (1996). See also D. Ertas and Y. Kantor, Phys. Rev. **E53**, 846 (1996) for another aspect of this problem, and S. Wolfling and Y. Kantor, Eur. Phys. J. **B12**, 569 (1999) for other generalizations.